Classification of Shallot Disease Diagnosis Using Combination Back Propagation Neural Network with SMOTE

Bain Khusnul Khotimah[1], Muhammad Syarief[2], Budi Dwi Satoto[3], Eka Malasari Rahcman[4], Yeni Kustiyahningsih[5]
[1,2,3,4,5] Faculty of Engineering, University of Trunojoyo Madura
bain@trunojoyo.ac.id

Email: bain@trunojoyo.ac.id

Abstract. Damage to the shallot crop causes crop loss for farmers. Classification of pests and diseases of shallots is very important for farmers so that farmers can easily recognize pests and diseases of shallots. Back Propagation Neural Network method is used as Artificial Intelligence to classify onion plant pests. while many classes that occur are not balanced because they are dominant in certain types of diseases. So it requires a method to improve the replication of the amount of data in the class that is not balanced with SMOTE. The trial evaluation method used is the 10-fold cross validation method. The results of the 10-fold cross validation test show that the Neural Network method with data balance with SMOTE produces a fairly high average accuracy of 97.20%, better than the Neural Network method which produces an average accuracy of 96.60%.

Keywords: Back Propagation Neural Network, SMOTE, Cross Validation, pests and diseases

1 Introduction

Shallots (Allium ascalonicum Linn.) is one of the main spice commodities developed in Indonesia. Shallots are multi-purpose vegetables that are used as spices to complement cooking spices, ingredients for the food industry and are used as traditional medicine (Putrasamedja, 1996). Some of the problems in the production of shallots include the high intensity of disease attacks, the availability of quality seeds that are not sufficiently precise, and the availability of superior varieties that are resistant to disease. Meanwhile, the application of good cultivation techniques has not been carried out optimally, farmer institutions have not been able to support farming, the scale of business is relatively small due to the narrowness of land ownership and weak capital [1]. The onion farmers concluded that the number of diseases that attacked shallots was the main cause of the decline in shallot income. Early diagnosis of shallot plant disease provides a concept. Measure of Believe (MB) is a hypothetical belief that is influenced by symptoms and Measure of Disbelieve (MD) is a hypothetical disbelief that is influenced by symptoms. This method measures certainty or uncertainty in diagnosing disease by implementing a classification system.

One of the systems with artificial intelligence that has been successfully developed at this time is the machine learning process (computer). A system with artificial intelligence can be said to experience machine learning that imitates the workings of the human biological nervous system is an artificial neural network system (Artificial Neural Network). The process of diagnosing onion diseases for farmers has a very important role, not only a high level of accuracy, but also a high level of speed in diagnosing plant diseases. Artificial Neural Network (ANN) is an artificial intelligence system that is able to analyze complex data and has an undefined or non-linear relationship between variables [9]. In previous studies, this ANN method has been widely applied in classification to diagnose plant diseases, such as corn with an accuracy rate of 95.08%. The dataset used in this study was obtained from observations of onion diseases in agricultural land in Sumenep Regency.

In real data, there are many situations where the number of instances in one class is much less than the number of instances in another class. This phenomenon is known as the imbalance class problem, which affects the performance of the classification algorithm to decrease. Most of the conventional models assign the majority class to the data and ignore the minority class because of the skewness of the data. To overcome the imbalance class problem, one of the methods used is sampling. The sampling method modifies the distribution of data between the majority and minority classes in the training dataset to balance the amount of data for each class. At the data preprocessing stage, it was identified that the dataset used in this study has a very large imbalance class problem where data instances with large rating values (>5) are far fewer than data instances with small/medium rating values. So we need a preprocessing method to overcome the imbalance class problem. One of the over-sampling techniques that can be used is the Syntetic Minority Over Sampling Technique (SMOTE). SMOTE can duplicate data synthetically so that the problem of different data distribution can be overcome [19].

This study uses ANN to classify shallot diseases for consideration in decision making. To validate the effectiveness of the proposed method, two experimental scenarios were carried out: first, the ANN algorithm was directly used for modeling without considering class imbalance. Then for the second scenario, SMOTE over-sampling is used to increase the number of datasets to achieve a balanced dataset.

2 Research Method

2.1 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is an over-sampling method in which data in the minority class is reproduced using synthetic data derived from data replication in the minority class. Over-sampling in SMOTE takes an instance of the minority class and then looks for the k-nearest neighbor of each instance, then generates a synthetic instance instead of replicating the minority class instance. Therefore, it can avoid the problem of excessive overfitting [19]. The algorithm that works on the first SMOTE will take the value of the difference between the vector of features in the minority class and the nearest neighbor value of the minority class and then multiply that value by a random number between 0 to 1. Furthermore, the calculation results are added to the feature vector so that the new vector value results are obtained [30].

$Xnew = Xi + (Xi\hat{}- Xi) \times \delta$ (6)

with
Xi = vector of features in minority class
$Xi\hat{}$= k-nearest neighbors for Xi
δ = random number between 0 to 1

2.2 Neural network

The most popular neural network algorithm is Back Propagation, the Back Propagation algorithm performs learning on a multi-layer feed forward neural network consisting of three layers, namely: input layer, hidden layer, and output layer (Han & Kamber, 2006). In some Neural network diagrams, it is possible to have more than one hidden layer, although most contain only one hidden layer which is considered sufficient for various purposes (Larose, 2006). Back Propagation is a Neural network algorithm for classification that uses gradient descent, Back Propagation looks for a set of weights that can model the data so as to minimize the average squared distance between the predicted network class and the actual class label of the data tuples (Han & Kamber, 2006). Each training data observation is processed through the network; the output value is generated from the input node. This output value is then compared with the actual value of the target variable and the resulting error is calculated (Larose, 2006).

Back Propagation performs an iterative learning process that tries to minimize errors from classification.

BP adjusts the neuron weights in a backward direction based on the error value in the learning process (Kusrini & Luthfi, 2009). In each process, the relation weights in the network are modified to minimize the Mean Squared Error (MSE) value between the predicted value of the network and the actual value. Modification of network relations is carried out in a backward direction (Kusrini & Luthfi, 2009). The learning steps in the Back Propagation method are as follows (Myatt, 2007):

1. Initialize all network weights randomly between -1.0 to 1.0)
2. Calculate the current input value and network weight using the formula:
   $I_j = \Sigma i\ w_{ij} O_i + \theta_j$
   With:
   O_i = Output node i from the previous layer
   w_{ij} = relation weight from node i in the previous layer *i* to node *j*
   θ_j= bias (as a limiter)
3. The output result is considered as the input from the second step, then generates the output for the node using the sigmoid activity function = 1 1+
4. Calculate the error value between the predicted value and the actual value using the formula:
   $Error_i = output_i * (1 - Outputi\ ) * (Target_i - Output_i)$
   With:
   $Output_i$ = Actual output of node i
   $Target_i$ = known target value in training data
5. After the error value is calculated, then it is reversed to the previous layer (Back Propagation). To calculate the error value in the hidden layer, use the formula.
   $Errori = Output_i * (1 - Output_i\ ) * \sum Error_j W_{ij}$

   With:
   $Error_i$ = Errors generated from hidden nodes *i*
   $Output_i$ = The output value of the hidden node *i*
   $Error_j$ = Error resulting from node *j* connected to output *i*

3 RESEARCH METHOD

Hybrid SMOTE and BPNN methods are implement unbalanced data class handling with SMOTE on training data with performing calculations to generate artificial (synthetic) data. So, randomly partition the dataset into 10 parts with a 10-fold cross validation scheme. Comparing the classification performance without and with the application of SMOTE on the ANN classification. Then for the second scenario, SMOTE over-sampling is used to increase the number of datasets to achieve a balanced dataset. This over-sampling technique was chosen because it has been widely applied to class imbalance problems.

Table 1. List of Features in the form of Symptoms of Shallot Attack

Feature	Explanation of symptoms
F1	White spots on leaves
F2	Leaf spots are white or gray
F3	The purple leaf spots are slightly reddish
F4	Green patches on uneven leaves
F5	Purplish red spots on the leaves
F6	Light brown leaf spots
F7	Shiny white leaf spots
F8	Purplish white leaf spots
F9	Yellow leaves
F10	Withered
F11	Burnt chocolate leaves
F12	Pale green or yellow striped leaves
F13	Dry and hollow leaves
F14	Small onion bulbs
F15	Stem neck cut
F16	Yellow leaves
F17	Twisted leaves wither and easy to pull out
F18	Bulbs rot from the tip
F19	Leaves spread underground
F20	Curved spots on white leaves
F21	Colored spots resembling a reddish-purple ring
F22	The tips of the leaves dry up even to the point of breaking
F23	Dried tubers are dark
F24	Leaf base shrinks
F25	Half of it is pink
F26	Plants grow stunted
F27	Small leaves
F28	Small Leaf Color
F29	Selected leaf growth
F30	Spots on the tips of the leaves

Table 1. describes the specific gravity of the symptoms that show damage to the shallots. While the identification of the type of disease and treatment solutions are in accordance with Table 2.

Table 2. List of labels in the form of onion plant diseases

No.	Types of Pests and Diseases
1	Leaf-slicing fly
2	Onion caterpillar pests
3	Pests Trips
4	Soil caterpillar
5	Wilt disease
6	Purple spot disease
7	Anthracnose disease
8	Mosaic Virus Disease
9	Leaf spot disease

4 RESULTS AND DISCUSSIONS

The determining the amount of training cycles carried out several experiments to get the best number of training cycles. In this experiment, the value of training cycles from 100 – 1000 with hidden layer by default in This experiment uses the value of parameters 0.3 learning rate and 0.2 for momentum. Results that obtained as in the Table 3.

Table 3. Determination of the best epoch

Training cycles (Epoch)	Accuracy	smote
100	93,01	93,5
200	95,00	95,6
300	95,00	96,71
400	96,64	97,65
500	95,75	95,92
600	93,54	93,53
700	95,27	97,66
800	95,32	96,71
900	96,87	97,59
1000	95,56	95,03

then the learning rate value is determined by doing a trial by entering a value with a range of 0.1 to 0.9. The value of training cycles is determined the same as from the previous experiment, which is 800, while 0.4 is used for the momentum value. The following are the results of the experiments that have been carried out to determine the learning rate value as shown in table 3 below: Table 3. Experiments for determining the learning rate as shown in Table 4.

Table 4. Experiments in determining the value of learning rate

Learning Rate	Accuracy	smote
0.1	93.07	93.52
0.2	95.04	95.63
0.3	95.13	96.72
0.4	96.26	96.87
0.5	97.52	95.76
0.6	93.57	93.53
0.7	95.69	95.68
0.8	96.72	96.71
0.9	97.07	97.60
1	95.25	95.72

The momentum value is determined by how to do by trial Enter values in the range 0.1 to with 0.9. Since, value of training cycles 800 and learning rate 0.9 is selected based on previous experiment. The following is the results of the experiments that have been carried out for the determination of momentum values such as in tabel 5 below:

Table 5. Experiments for determining the value of momentum

Momentum	Accuracy	smote
0.1	94.09	93.59
0.2	95.80	95.68
0.3	95.80	96.76
0.4	96.08	97.00
0.5	96.24	95.04
0.6	96.53	93.57
0.7	96.61	95.65
0.8	97.76	96.73
0.9	97.23	97.50
1	95.10	95.87

The combination of BPNN and SMOTE, with changes in several parameters produces a higher accuracy than BPNN. In this case, the distribution of the data distribution can be balanced because the class with a smaller number of samples (minority class) will be multiplied by synthetic data made by SMOTE. so that the test results are as if for a balanced class

References

1. Aldo, Dasril dan Sapta, Eka. “Sistem Pakar Diagnosis Hama dan Penyakit Bawang Merah Menggunakan Metode Dempster Shafer”, Jurnal Sistem Komputer. Vol. 9, No. 2 : 85-93, 2019.
2. G. Kesavaraj and S. Sukumaran, “A study on classification techniques in data mining,” in 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013, 2013, doi: 10.1109/ICCCNT.2013.6726842.
3. [6] M. M. Saritas and A. Yasar, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification,” Int. J. Intell. Syst. Appl. Eng., vol. 7, no. 2, pp. 88–91, Jun. 2019, doi: 10.18201//ijisae.2019252786.
4. Coupled Water Tank,” Bull. Comput. Sci. Electr. Eng., vol. 1, no. 1, pp. 12–18, 2020, doi: 10.25008/bcsee.v1i1.4.
5. H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
6. Y. Sun, A. K. C. Wong, and M. S. Kamel, “Classification of imbalanced data: A review,” Int. J. Pattern Recognit. Artif. Intell., vol. 23, no. 4, pp. 687–719, Jun. 2009, doi: 10.1142/S0218001409007326.
7. Chowdhary K R 2020 Fundamentals of Artificial Intelligence (New Delhi: Springer Nature)
8. N. B. D. Timea Bezdan, "Convolutional Neural Network Layers And Architectures," Data Science & Digital Broadcasting Systems, Vol. 10, Pp. 445-451, 2019.
9. W S Simamora, R S Lubis, and E M Zamzami, A Classification: using Back Propagation Neural Network Algorithm to Identify Cataract Disease, Journal of Physics: Conference Series, Volume 1566, 4th International Conference on Computing and Applied Informatics 2019 (ICCAI 2019) 26-27 November 2019, Medan, Indonesia
10. R A Dilruba, N Chowdhury, F F Liza, and C K Karmakar 2006 Data Pattern Recognition using Neural Network with Back-Propagation Training Proc. Int. Conf. on Electrical and Computer Engineering Dhaka, Bangladesh, 2006, p. 451–455
11. Alaeldin S and Yun Z 2015 A Review on Back-Propagation Neural Networks in the Application of Remote Sensing Image Classification J. Earth Sci. Eng 5 1
12. Hussein Attya Lafta, ZainabFalah Hasan, NoorKadhim Ayoob, Classification of medical datasets using back propagation neural network powered by genetic-based features elector, International Journal of Electrical and Computer Engineering (IJECE)Vol.9, No.2, April, 2019, pp. 1379~1384